Prediction of Wildfire Cause

Mohammad Bhuiyan
bhui8420@mylaurier.ca

Abstract

The world was ravaged by wildfire in the last decades. Over 52,000 wildfires recorded in the USA in 2018 alone [1]. We observed devastating fire in Australia in 2019 that claims nearly 3 billion wildlife [2] and burned down 18 million hectares of land [3]. Wildfire caused devastating ecological and economic impact and loss of lives. Better wildfire management and loss mitigation could be possible by predicting cause of wildfire. This project attempts to predict cause of wildfire from data of wildfire that occurred in the USA over the period of 24 years [5]. The model designed in this project was able to predict cause of wildfire with 70% accuracy. The accuracy of 92% was achieved predicting Arson in California.

1. Description of Applied Problem

Wildfire management becomes a major concern in today's world. With the global warming, the world observed devastating infernos in recent years. Apart from immediate damage to ecology and economy, wildfire leave lasting effect on global warming and environmental pollution. Large amount of CO_2 and other greenhouse gases released in the atmosphere during the wildfire contribute to global warming which in turn increase risk of wildfire. The global warming is worsening with this climate feedback loop [4].

There is always resource shortage to fight large scale wildfire. Wildfire could be managed efficiently and even mitigate significantly if there were prediction tool to know the cause of upcoming risk of wildfire. Machine learning can help make such predictions by analyzing historical data with features that influence the cause of wildfire.

2. Description of Available Data

Wildfire record of the national Fire Program Analysis (FPA) system of the USA is 24 years of geo-referenced wildfire records. The data was collected for wildfires that occurred in the USA from 1992 to 2005. The data made available for study in Kaggle as "1.88 Million US Wildfires" and could be downloaded as SQLite database. The database represents a total of 140 million acres burned during the 24-year period [5]. Out of tons of information stored in the database, dataset relevant to this study are listed below [Table 1]:

Table 1: Description of Dataset [5]

- Fires: Table including wildfire data for the period of 1992-2015 compiled from US federal, state, and local reporting systems.
- FIRE_YEAR = Calendar year in which the fire was discovered or confirmed to exist.
- DISCOVERY_DATE = Date on which the fire was discovered or confirmed to exist.
- DISCOVERY_DOY = Day of year on which the fire was discovered or confirmed to exist.
- DISCOVERY_TIME = Time of day that the fire was discovered or confirmed to exist.
- STAT*CAUSE*CODE = Code for the (statistical) cause of the fire.
- STAT*CAUSE*DESCR = Description of the (statistical) cause of the fire.
- CONT_DATE = Date on which the fire was declared contained or otherwise controlled (mm/dd/yyyy where mm=month, dd=day, and yyyy=year).
- CONT_DOY = Day of year on which the fire was declared contained or otherwise controlled.
- CONT_TIME = Time of day that the fire was declared contained or otherwise controlled (hhmm where hh=hour, mm=minutes).
- FIRE_SIZE = Estimate of acres within the final perimeter of the fire.
- FIRE*SIZE*CLASS = Code for fire size based on the number of acres within the final fire perimeter expenditures (A=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1000 to 4999 acres, and G=5000+ acres).
- LATITUDE = Latitude (NAD83) for point location of the fire (decimal degrees).
- LONGITUDE = Longitude (NAD83) for point location of the fire (decimal degrees).
- STATE = Two-letter alphabetic code for the state in which the fire burned (or originated), based on the nominal designation in the fire report.
- COUNTY = County, or equivalent, in which the fire burned (or originated), based on nominal designation in the fire report.

3. Analysis and Visualization Techniques

3.1. Pre-Processing

The data was in SQLite database format. Preliminary data analysis was done with “DB Browser for SQLite Version 3.12.2” with the advantage of SQL statements. Subsequently, we used python libraries like sqlite3, pandas and NumPy for exploratory purposes. We found our data with required features in ‘Fires’ table. Dates were in Julian format and later it was converted to Gregorian format from which month and day-of-week were extracted. There

were 13 causes of wildfire listed in the database which were categorized into four groups for better correlation with other features. All the features were converted to numerical data.

Table 2: Initial Data from the table Fires

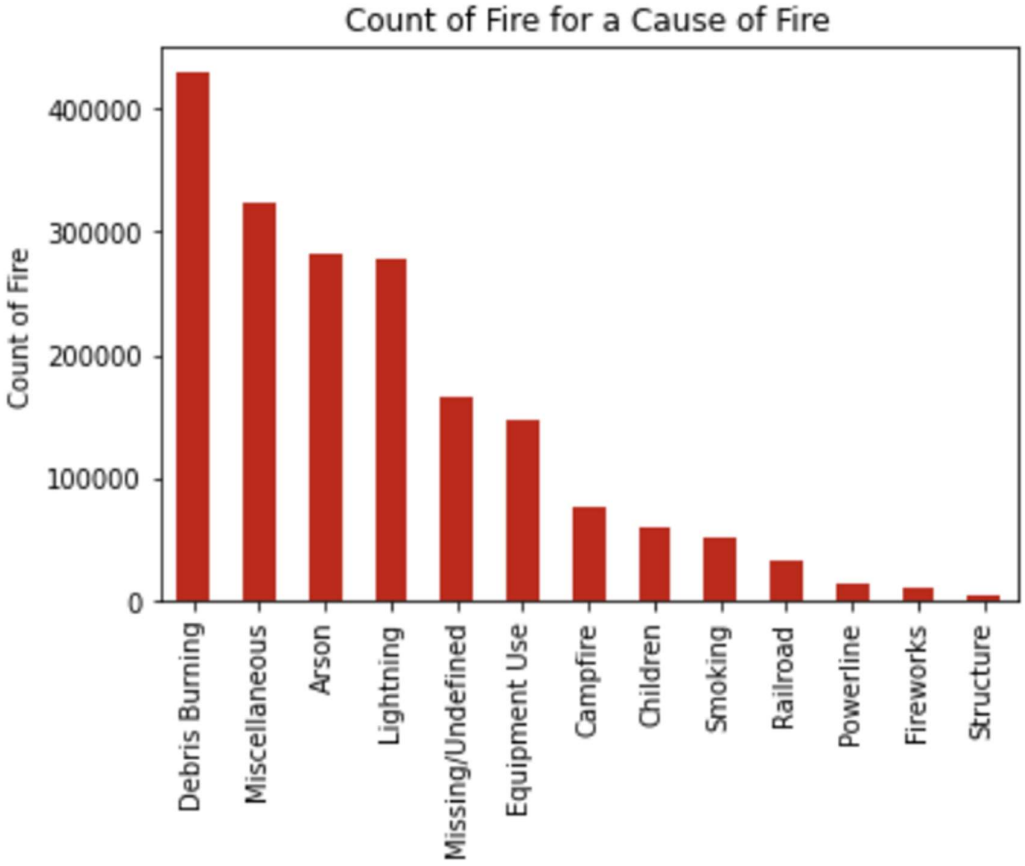
FIRE_YEAR	STAT_CAUSE_DESCR	LATITUDE	LONGITUDE	STATE	DISCOVERY_DATE	FIRE_SIZE
2005	Miscellaneous	40.036944	-121.005833	CA	2453403.5	0.1
2004	Lightning	38.933056	-120.404444	CA	2453137.5	0.25
2004	DebrisBurning	38.984167	-120.735556	CA	2453156.5	10
2004	Lightning	38.559167	-119.913333	CA	2453184.5	0.1
2004	Lightning	38.559167	-119.933056	CA	2453184.5	0.1
–	–	–	–	–	–	–

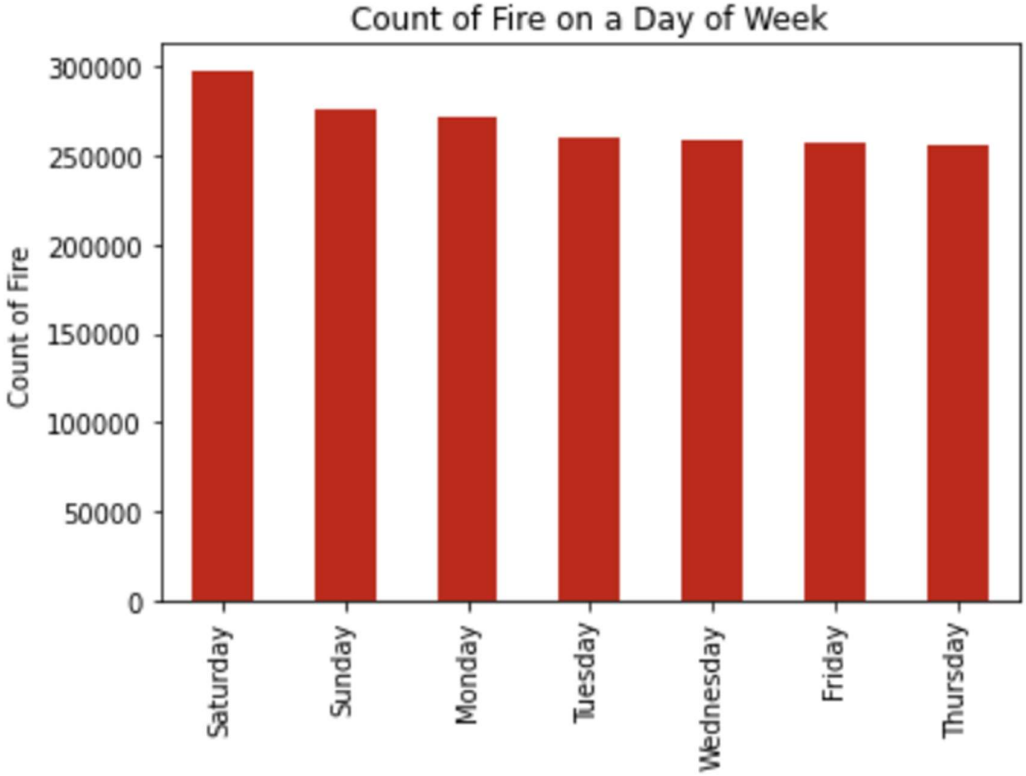
Table 3: Processed data of the table Fires

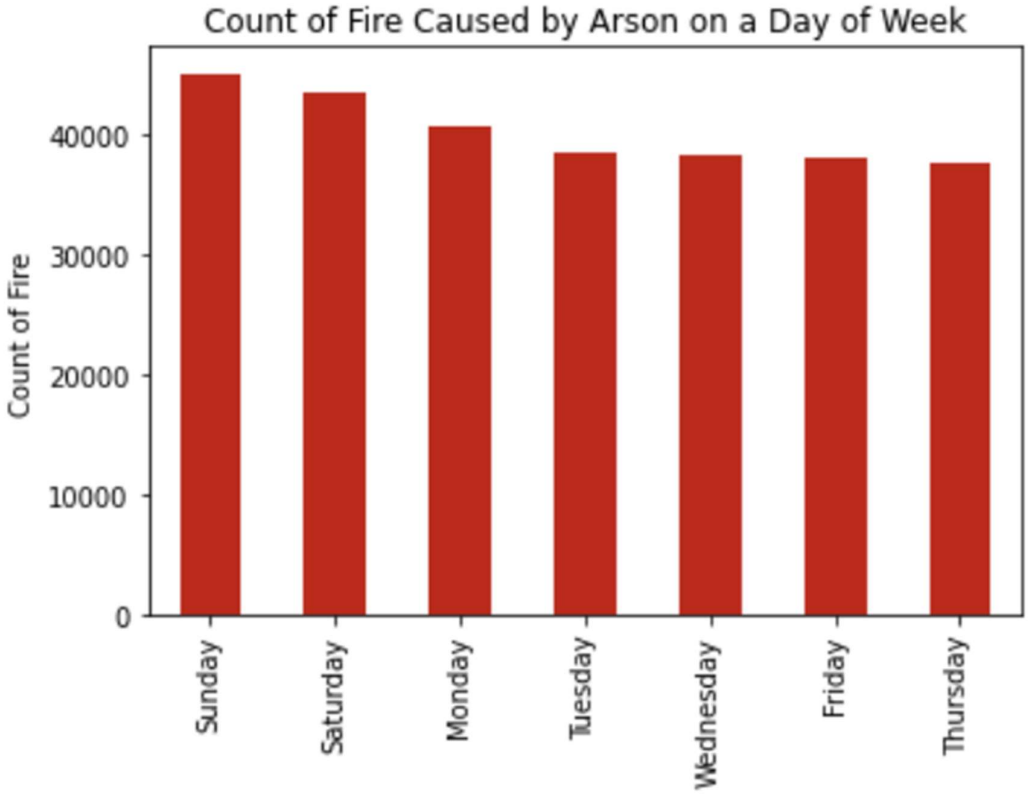
FIRE_YEAR	STAT_CAUSE_DESCR	LATITUDE	LONGITUDE	STATE	DAY_OF_WEEK	MONTH	FIRE_SIZE
2005	7	40.036944	-121.005833	4	6	2	0.1
2004	6	38.933056	-120.404444	4	6	5	0.25
2004	3	38.984167	-120.735556	4	1	5	0.1
2004	6	38.559167	-119.913333	4	1	6	0.1
2004	6	38.559167	-119.933056	4	1	6	0.1
–	–	–	–	–	–	–	

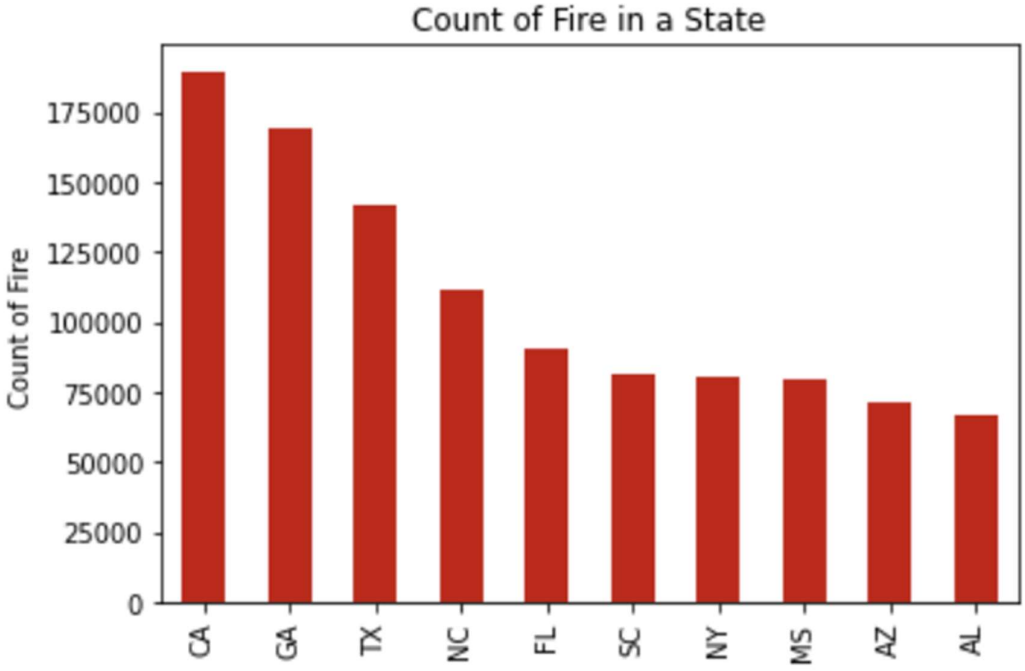
3.2. Analysis

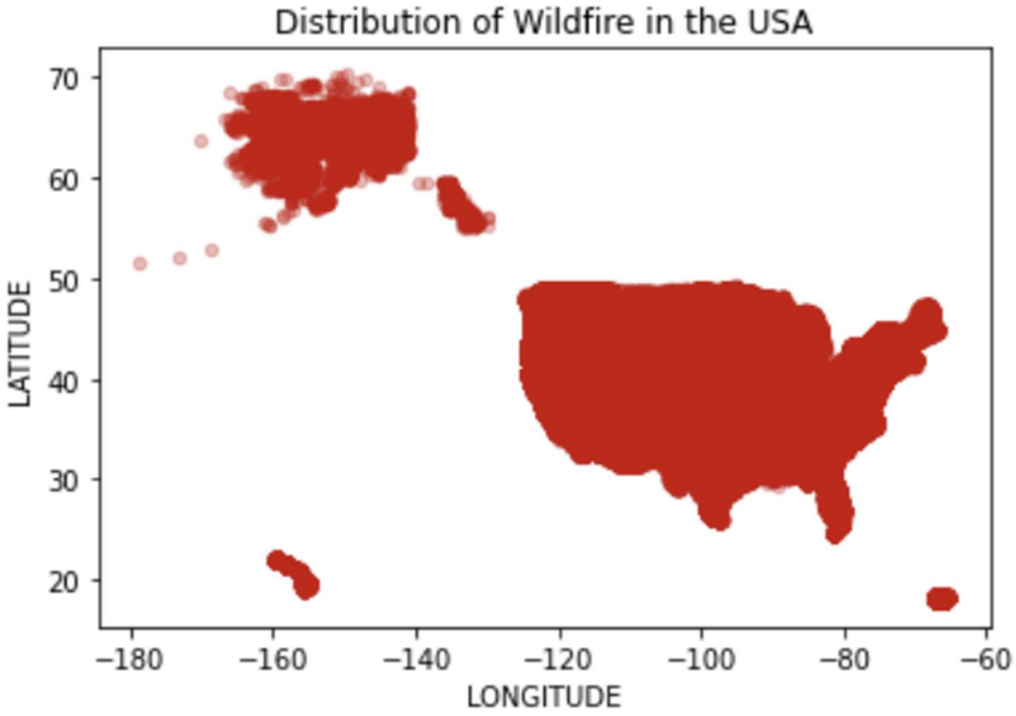
Following features were extracted from the table ‘Fires’ to design the model: FIRE_YEAR, STAT_CAUSE_DESCR, LATITUDE, LONGITUDE, STATE, DISCOVERY_DATE, FIRE_SIZE. Two new features, day-of-week and month were extracted from the feature discovery_date and the discovery_date was dropped later. There were 13 causes of wildfire in the database. Upward trend of fire incident observed during the weekend and this upward trend had strong relation with Arson related fire. Fire count varied significantly from one state to another. Top three states that contributed most to the fire counts were California, Georgia and Texas.


Count of Fire for a Cause of Fire
Count of Fire
400000
300000
200000
100000
0
Debris Burning
Miscellaneous
Arson
Lightning
Missing/Undefined
Equipment Use
Campfire
Children
Smoking
Railroad
Powerline
Fireworks
Structure


Count of Fire on a Day of Week
Count of Fire
300000
250000
200000
150000
100000
50000
0
Saturday
Sunday
Monday
Tuesday
Wednesday
Friday
Thursday

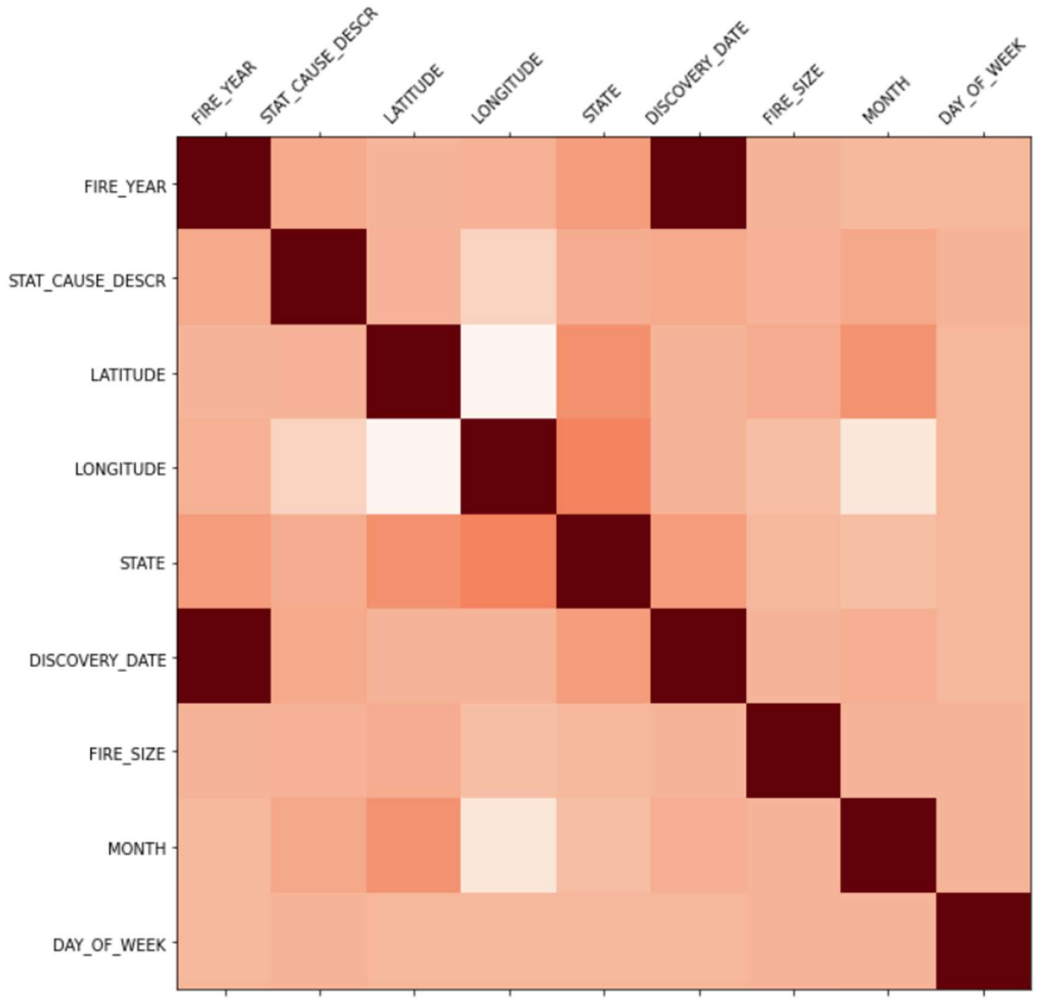

Count of Fire Caused by Arson on a Day of Week
Count of Fire
40000
30000
20000
10000
0
Sunday
Saturday
Monday
Tuesday
Wednesday
Friday
Thursday


Count of Fire in a State
Count of Fire
175000
150000
125000
100000
75000
50000
25000
0
CA
GA
TX
NC
FL
SC
NY
MS
AZ
AL



3.3. Visualization

Data visualization is important to find correlation among large number of variables and interpret the relationships meaningfully. Confusion Matrix, Correlation Matrix, Histogram and Scatter plots were used for data visualization. Correlation matrix was created to understand how the features were related.



strong correlations are in dark color and there is no correlation if the color is white

Confusion metrics were drawn after analyzing the data with various machine learning algorithm to visualize the accuracy of the model.

4. Prediction Model

Supervised classification technique, Random Forest Classification was selected after experimenting with other techniques like Decision Tree, SVM and Neural Network. The dataset was divided into 25% and 75% for training and testing purposes and STAT_CAUSE_DESCR was labelled. We achieved 58% accuracy in predicting 13 classes using Random Forest Classification.

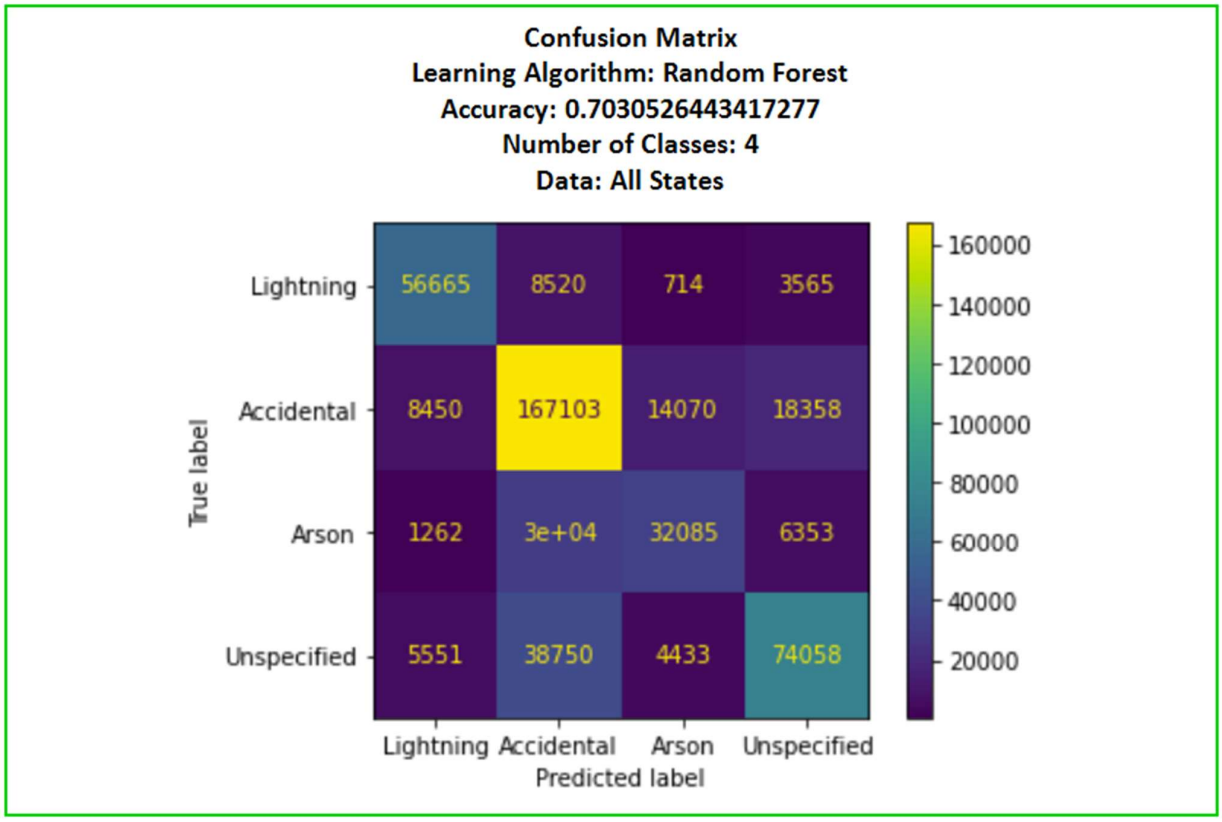
Confusion Matrix
Learning Algorithm: Random Forest
Accuracy: 0.5812616859207389
Number of Classes: 13
Data: All States

37993	628	845	18153	1972	123	1619	6945	968	64	249	309	12
1123	6172	209	3850	801	49	3390	2671	401	23	110	207	9
2110	288	2305	5213	1081	119	749	2361	262	26	382	253	24
14464	1504	1241	73072	3296	160	2590	9121	843	173	526	585	45
3058	575	542	9789	10800	96	2766	7206	687	146	927	333	24
290	65	100	344	134	1207	315	336	38	12	5	19	5
1181	986	135	2742	1025	61	58949	3248	758	65	184	126	4
6327	1757	941	16173	5489	226	4930	41960	1514	204	686	690	32
532	222	75	1068	421	21	2089	1582	35761	15	12	64	1
245	58	29	1066	373	29	392	941	83	355	12	18	5
472	179	108	1510	790	6	776	565	66	10	3870	53	0
1382	600	348	4551	967	44	962	3285	233	27	180	776	3
127	32	57	373	76	23	56	166	11	9	5	9	41

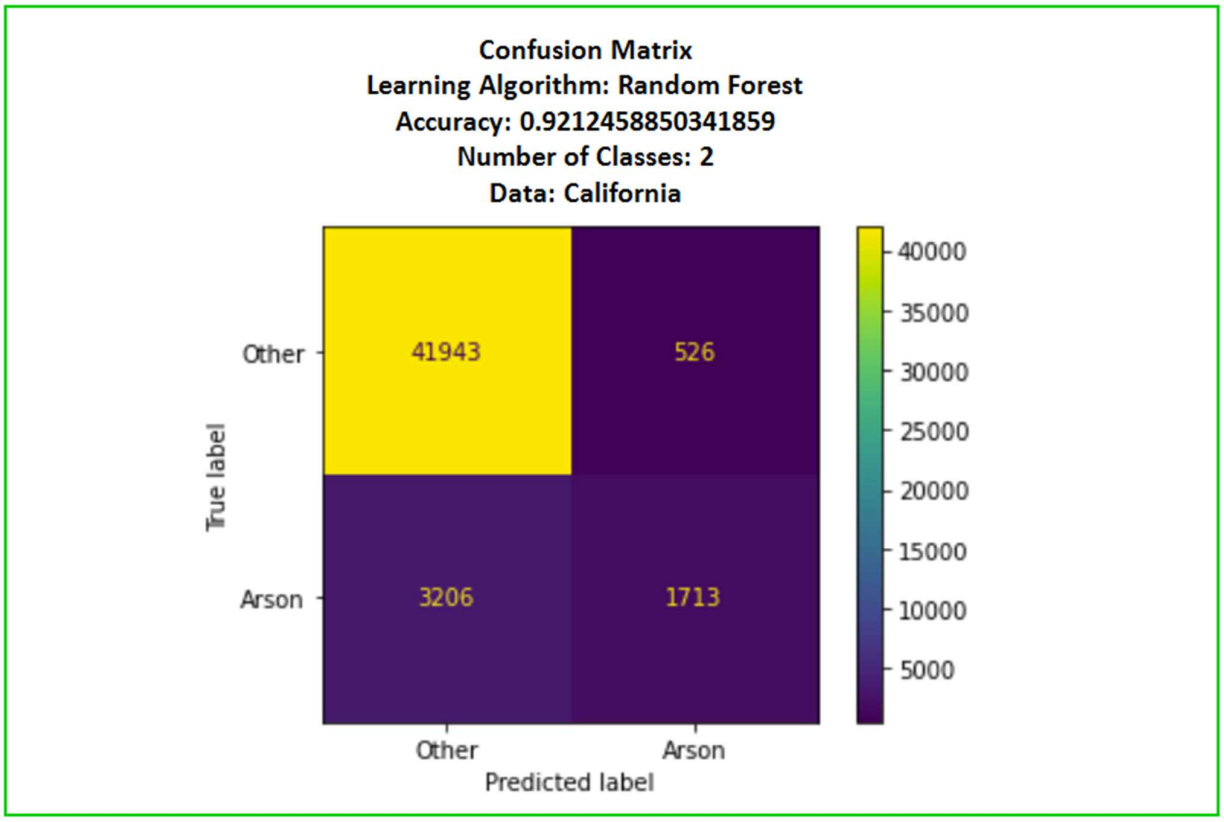
We grouped 13 causes of wildfire into four groups to achieve higher accuracy. The causes were grouped as below:

```
Lightning = ['Lightning']
Accidental = ['Structure', 'Fireworks', 'Powerline', 'Railroad', 'Smoking', 'Children', 'Campfire', 'Equipment
    Use', 'Debris Burning']
Arson = ['Arson']
Unspecified = ['Missing/Undefined', 'Miscellaneous']
```

After reducing the number of classes into four, we were able to achieve 70% accuracy which was significant improvement.



We observed some correlation between fire count and weekend. We analyzed the data against each cause and found arson was the main contributing factor of increased fire incident during the weekend. Based on this observation we categorized all the causes into two groups, Arson and Other. Then we run the learning algorithm Random Forest Classification against Wildfire records of California where highest number of wildfires were reported. This time we achieved 92% accuracy predicting Arson.



Main features of the model are summarized in below table:

Table 4: Model Summary

Learning Algorithm: Random Forest Classification				
Number of Classes	Range of Data	Hyper-parameters	Accuracy	MSE
13	All States	n_estimators = 50	58%	7.91
4	All States	n_estimators = 50	70%	0.83
2	California	n_estimators = 200	92%	0.08

MSE: Mean Squared Error

5. Conclusion

Wildfire management is a big concern in today's world. Adequate resource allocation and quick resource relocation are utmost important in containment of colossal disaster. For any large-scale wildfire, resources would never be close to enough. Our only hope is to early detection and mobilize the resources as quickly as possible. This predictive model can help us predict cause of wildfire with 70% accuracy in general. Significant correlation was observed between weekend and arson as a cause of fire. The model attempts to predict arson

from the wildfire data of California, the top fire prone state, and it was able to predict the arson with 92% accuracy. The model could be extended for other States with some fine tuning. This model would be great help to prepare ourselves for ensuing disaster.

6. References

[1] URL https://www.iii.org/fact-statistic/facts-statistics-wildfires

[2] URL https://www.cnn.com/2020/07/28/asia/australia-fires-wildlife-report-scli-intl-scn/index.html

[3] URL https://towardsdatascience.com/leveraging-machine-learning-to-predict-wildfires-contributing-to-the-united-nations-sustainable-a10c5044dcae

[4] URL https://towardsdatascience.com/predicting-california-wildfire-size-with-neural-networks-building-a-machine-learning-project-from-db0e57dce4c9

[5] URL https://www.kaggle.com/rtatman/188-million-us-wildfires

[6] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In icml, volume 99, pages 124-133, 1999.